HTN Method Selection Guided by Multi-Objective Reinforcement Learning and Dynamic Preferences

Anderson G. A. C. Filho[1] **, Anderson R. Tavares**[1] **, Lucas N. Alegre**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS, Brasil

{anderson.goncalves,artavares,lnalegre}@inf.ufrgs.br

Abstract. *This work proposes a hybrid framework in which Hierarchical Task Network (HTN) method selection is modeled as a temporally extended hierarchical decision, enabling the selected method to receive credit according to multiple objectives, each defined by a different reward function. This work aims to investigate whether the integration of symbolic planning and multi-objective reinforcement learning can provide adaptive agents while preserving interpretability, designer control, and explicit semantic constraints. The planned evaluation compares the proposed approach against traditional HTN method-selection strategies and scalar reinforcement learning baselines, considering adaptation to different preferences, generalization, semantic validity, and inference cost.*

1. Introduction

Symbolic planning frameworks are widely adopted in digital games and robotics because they provide interpretable representations of agent behavior and allow developers to encode explicit domain knowledge and constraints [Ghallab et al. 2004, Millington and Funge 2019]. Among these approaches, Hierarchical Task Network (HTN) planning decomposes compound tasks C_t, into progressively simpler subtasks through domain-defined methods, until primitive tasks P in $\mathcal{P}$, corresponding to executable actions in the environment are obtained [Erol et al. 1994, Nau et al. 2003, Georgievski and Aiello 2014].

A central challenge in HTN-based agents is *method selection*: when multiple decomposition methods are applicable, the agent must determine which alternative is preferable under the current context. Previous work has investigated both reinforcement-learning-based method selection [Hogg et al. 2010, Bahrami 2025] and multi-objective HTN planning [Li et al. 2024]. However, in practical applications, "good" behavior is rarely characterized by a single scalar objective. Designers may simultaneously care about criteria such as success, cost, risk, time, resource usage, or play style, whose relative importance may vary across users, missions, or over time.

Multi-objective reinforcement learning (MORL) provides a framework for sequential decision making under vector-valued rewards and potentially conflicting objectives [Roijers et al. 2013, Hayes et al. 2022]. In particular, preference-conditioned approaches allow the behavior of a policy to change according to different objective weights without requiring a separate policy for every preference configuration [Abels et al. 2019]. This work investigates how such preference-dependent decision making can be integrated with HTN planning while preserving the symbolic constraints imposed by the planner.

We model HTN method selection as a temporally extended hierarchical decision. A selected method may decompose into a sequence of subtasks whose execution spans multiple primitive environment steps. Consequently, the selected method receives credit according to the vector return accumulated until the next hierarchical decision, episode termination, or replanning event, consistently with semi-Markov formulations of temporal abstraction [Sutton et al. 1999]. Importantly, an HTN method is not treated as an RL option; it remains a symbolic decomposition rule defined by the domain, while reinforcement learning is used to evaluate and select among applicable methods.

We propose a hybrid framework for HTN method selection guided by MORL and dynamic preferences. The HTN component determines which decomposition methods are semantically applicable, whereas the learned policy selects among these alternatives according to their estimated multi-objective returns and the current preference vector.

At this stage, we report the methodological formulation and evaluation protocol; the experimental results are not yet available.

2. Related Work

HTN planning decomposes tasks through domain-defined methods and constraints [Erol et al. 1994, Georgievski and Aiello 2014]. Reinforcement learning has previously been used to guide HTN method selection. Hogg et al. [Hogg et al. 2010] learn scalar task–method values, while Bahrami [Bahrami 2025] investigates Deep Q-Learning for selecting among applicable methods. Multi-objective HTN planning has instead used dominance-based search to retain plans representing alternative objective tradeoffs [Li et al. 2024]. In contrast, our framework learns vector-valued return estimates for HTN-applicable methods and conditions runtime method selection on a changing preference vector.

3. Proposed Methodology

Given a compound task C_t and symbolic state S_t, the HTN planner first filters the decomposition methods according to their preconditions. A preference-conditioned MORL policy then evaluates only the applicable methods using a normalized preference vector $\mathbf{w}_t$, where $w_{t,i} \in [0, 1]$ and $\sum_i w_{t,i} = 1$. The components of $\mathbf{w}_t$ represent the relative importance of objectives such as task progress, survival, resource conservation, and execution time. Method selection is defined as

$$M_t^* = \arg \max_{M \in \mathcal{M}_{\text{valid}}(C_t, S_t)} \mathbf{w}_t^\top \mathbf{Q}_\theta(S_t, M; C_t, \mathbf{w}_t), \tag{1}$$

where $\mathbf{Q}_\theta$ estimates the vector-valued return of an applicable method (in the set $\mathcal{M}_{\text{valid}}(C_t, S_t)$), and $\mathbf{w}_t^\top \mathbf{Q}_\theta$ scalarizes it according to the current preferences. For illustration, consider a fixed decision context with two objectives (e.g., progress and survival), and fixed vector estimates $\mathbf{Q}_\theta(m_1) = [0.8, 0.2]$ and $\mathbf{Q}_\theta(m_2) = [0.5, 0.5]$. Then $\mathbf{w}_t = [0.8, 0.2]$ selects m_1, whereas $\mathbf{w}_t = [0.2, 0.8]$ selects m_2.

Because a method may generate subtasks spanning multiple primitive steps, its selection is modeled as a temporally extended decision. If the j-th hierarchical decision lasts k_j steps, its vector return is $\mathbf{R}_j^M = \sum_{i=0}^{k_j-1} \gamma^i \mathbf{r}_{t_j+i}$.

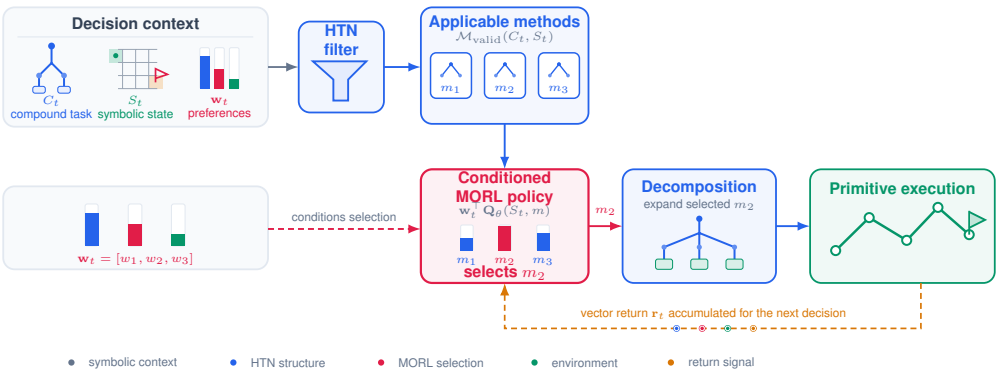


Figure 1. Visual summary of the proposed decision cycle. Symbolic context and preferences constrain and rank applicable HTN methods; the selected method is decomposed and executed, and its multi-objective return is assigned at the next hierarchical decision.

The framework will be evaluated in a controlled game-like environment with multiple objectives and dynamic preferences. It will be compared against (i) first-applicable HTN selection, (ii) fixed heuristic HTN selection, (iii) scalar RL for task–method values [Hogg et al. 2010], and (iv) contextual scalar RL using the same HTN applicability mask. Dynamic preferences will be induced by switching $\mathbf{w}_t$ at predefined hierarchical decisions. We will measure adaptation through post-switch scalarized return and changes in method selection, generalization on held-out preference vectors and initial states, semantic validity through HTN-applicable selections and constraint-satisfying decompositions, and inference cost through decision latency.

4. Conclusion

This work defines a framework for preference-conditioned MORL over HTN-valid decomposition methods. HTN planning preserves symbolic constraints, whereas MORL adapts method selection to the current state and objective priorities. By assigning the accumulated vector return to the selected method, the formulation links high-level symbolic choices to their multi-objective consequences. The planned evaluation will assess adaptation, generalization, semantic validity, and online decision cost against fixed HTN and scalar RL baselines.

Acknowledgments

This research was supported in part by CAPES, Finance Code 001, and partially funded by FAPERGS (grant 26/2551-9000125-0). This work was supported by Kunumi Institute. The authors thank the institution for its financial support and commitment to advancing scientific research.

References

Abels, A., Roijers, D. M., Lenaerts, T., Nowé, A., and Steckelmacher, D. (2019). Dynamic weights in multi-objective deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 11–20.

Bahrami, S. (2025). Enhancing htn planning with deep reinforcement learning for method selection. Master's thesis, University of Stuttgart.

Erol, K., Hendler, J., and Nau, D. S. (1994). Htn planning: Complexity and expressivity. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*. AAAI Press.

Georgievski, I. and Aiello, M. (2014). An overview of hierarchical task network planning. *arXiv preprint arXiv:1403.7426*.

Ghallab, M., Nau, D., and Traverso, P. (2004). *Automated Planning: Theory and Practice*. Elsevier, San Francisco.

Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L. M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A. A., Mannion, P., Nowé, A., Ramos, G., Restelli, M., Vamplew, P., and Roijers, D. M. (2022). A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1).

Hogg, C., Kuter, U., and Muñoz-Avila, H. (2010). Learning methods to generate good plans: Integrating htn learning and reinforcement learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, pages 1530–1535.

Li, M., Liu, X., Jiang, G., and Liu, W. (2024). A novel hierarchical task network planning approach for multi-objective optimization. *Expert Systems with Applications*, 251:124058.

Millington, I. and Funge, J. (2019). *Artificial Intelligence for Games*. CRC Press, Boca Raton, 3 edition.

Nau, D., Au, T.-C., Ilghami, O., Kuter, U., Murdock, J. W., Wu, D., and Yaman, F. (2003). Shop2: An htn planning system. *Journal of Artificial Intelligence Research*, 20:379–404.

Roijers, D. M., Vamplew, P., Whiteson, S., and Dazeley, R. (2013). A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113.

Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211.